DOCUMENT RESUME

ED 264 302 TM 860 010

| | |
|---|---|
| AUTHOR | Abbott, Muriel M. |
| TITLE | Theoretical Considerations in the Measurement of the English-Language Proficiency of Limited-English-Proficient Students. |
| PUB DATE | 1 Apr 85 |
| NOTE | 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL., March 31-April 4, 1985). |
| PUB TYPE | Speeches/Conference Papers (150) -- Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | Bilingual Education Programs; Elementary Secondary Education; English (Second Language); *Language Proficiency; Language Tests; *Limited English Speaking; *Measurement Techniques; *Program Evaluation; Student Evaluation; *Student Placement; Test Construction |
| IDENTIFIERS | *Language Assessment Battery; New York City Board of Education |

ABSTRACT

Limited-English-proficient (LEP) students who are not sufficiently proficient in English to participate effectively in an English-speaking instructional environment must be identified for placement in an appropriate educational program. Their progress through an instructional program that is designed to improve their acquisition of English language skills must also be measured both for purposes of exiting from the program and for program evaluation. Both placement and evaluation require a measure that assesses different levels of English-language proficiency. Attention is directed to the concept of language proficiency, how best to measure it with relevance to curriculum, appropriate levels of difficulty, and a meaningful frame of reference for the interpretation of scores for different populations. Attention is also directed to issues concerned with constructing an instrument that can serve both placement and evaluation purposes; that is, an instrument that can differentiate efficiently between English-proficient and LEP students yet, at the same time, provide efficient measurement of the progress of the lower-scoring LEP students. These issues are discussed with reference to the New York City Language Assessment Battery (LAB-1982). (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED264302

# Theoretical Considerations in the Measurement of the English-Language Proficiency of Limited-English-Proficient Students

Muriel M. Abbott

New York City Board of Education

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Abbott, M.M.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, April 1, 1985.

TM 860 010

ABSTRACT

Theoretical Considerations in the Measurement of the English-language Proficiency of Limited-English-Proficient Students

Limited-English-proficient students who are not sufficiently proficient in English to participate effectively in an English-speaking instructional environment must be identified for placement in an appropriate educational program. Their progress through an instructional program that is designed to improve their acquisition of English language skills must also be measured both for purposes of exiting from the program and for program evaluation. Both placement and evaluation require a measure that assesses different levels of English-language proficiency. Attention is directed to the concept of language proficiency, how best to measure it with relevance to curriculum, appropriate levels of difficulty, and a meaningful frame of reference for the interpretation of scores for different populations. Attention is also directed to issues concerned with constructing an instrument that can serve both placement and evaluation purposes; that is, an instrument that can differentiate efficiently between English-proficient and limited-English-proficient students yet, at the same time, provide efficient measurement of the progress of the lower-scoring limited-English-proficient students. These issues are discussed with reference to the New York City Language Assessment Battery (LAB-1982).

# Theoretical Considerations in the Measurement of the English-Language Proficiency of Limited-English Proficient Students

Muriel M. Abbott

In order to develop a relevant and useful test, there must be a clear purpose for which the test results are required. The purpose in turn, determines the attribute(s) to be assessed and how these attributes should be measured. The primary purpose of the Language Assessment Battery (LAB) is to identify those students who are non-native speakers of English and who are not sufficiently proficient in English to participate effectively in an English-language instructional environment, the classroom. These students are eligible and entitled to special educational programs, either bilingual or English-as-a-second-language (ESL). The 1976 Language Assessment Battery was the first New York City test to serve this purpose and was developed to comply with the 1974 Lau v. Nichols[1] decision and the 1974 Aspira Consent Decree.[2] The 1982 Language Assessment Battery was also developed to comply with these mandates but represents a completely different approach to the measurement of language proficiency. The English version of the test was developed primarily to identify "entitled" students. A second purpose for the English version of the Language Assessment Battery was to measure gain in English proficiency or in the acquisition of English-language skills in order that student progress in an instructional program be assessed. It should be noted that the Aspira Consent Decree also mandated that for Hispanic students language of dominance (English or Spanish) be identified. To comply with this mandate the Spanish version of LAB was constructed. The Spanish version is not a translation of the English version but parallels it and was developed concurrently with it.

The two purposes, placement in program and measurement of gain, are related in that both are concerned with measuring differences in level of English-language proficiency. A placement test must be sensitive to different levels of English proficiency in order to provide a yardstick for entering into and exiting from an instructional program. An evaluation test must also be sensitive to different levels of English proficiency in order to permit measurement of gain as a result of an instructional program. To serve both purposes appropriately, the test must reflect curriculum content. Furthermore, a test designed to serve either purpose should be particularly sensitive at, and below the criterion cut-off for program eligibility. It was decided, therefore, that the same test could serve both purposes. However, because most tests are designed for English-proficient students, they have insufficient floor to be appropriate for measuring the performance of limited-English-proficient students. Therefore,the proposed test had to address the special needs of the limited-English-proficient group.

[1]Lau v. Nichols, 414 U.S. 563, 1974

[2]Aspira of New York, Inc., et.al. v. Board of Education of the City of New York, et.al., Memorandum 72 C.V. 4002, 1974

The purposes having been determined, the next task was to define the attribute to be measured: language proficiency.[3] This was defined as communicative competence, that is the ability to receive and convey information through the use of oral and written language. A review was made of the literature concerning the structure of language proficiency. The Partial Divisibility Hypothesis discussed by Oller[4] seemed to offer the most useful approach for test development. This hypothesis holds that a large amount of variance is common to all the tests measuring the many aspects of language proficiency as well as small amounts of variance unique to certain tests or combinations of tests. The Language Assessment Battery, therefore, was based on the belief that a single unitary factor underlies the many processing tasks involved in language proficiency but that there are also components unique to its various aspects. In the case of LAB for test construction purposes these components were identified as listening, reading, writing and speaking. This assumed that score interpretation for each of these areas separately would be worthwhile. Once the structure of language proficiency was hypothesized, it was necessary to determine how to elicit the behaviors that would reflect the various aspects that had been hypothesized. The task, then, was to relate language proficiency testing to language learning, especially the learning of English language skills in the classroom and to produce a test that would yield useful scores in each of the four subtest areas as well as a total "proficiency" score.

The next step was to determine in general how to approach the actual measurement of language proficiency. Again, the development of LAB was influenced by the work of Oller and Carroll who support an integrative, in preference to a discrete - point approach. The discrete point approach consists of taking apart language skills and measuring those individual segments separately. The integrative approach, however, combines segments into large units and measures the ability to make use of these larger language elements in the context of normal communication. A test constructed on this principle attempts to mirror normal language usage. Concomitent with this approach is the concept of "expectancy." Language in normal usage has a sequential organization. In a sequence of language elements (sounds, syllables, words, sentences) there is an expectancy of what will come next. Thus, the Language Assessment Battery was constructed on the premise that language proficiency is basically a single unitary factor but comprises reliably measurable separate elements that have curriculum relevance and that assessment of language proficiency should use to the greatest extent possible an integrative approach in measuring relevant curriculum objectives.

[3]The discussion that follows applies equally to the English and Spanish versions but for simplicity of explanation, reference is made only to the English version.

[4]Oller, John W. Jr. 1979. Language Tests at School. London, England: Longman Group Limited.

The selection of objectives to be measured resulted from a review of the literature as well as careful examination of New York City curriculum. Within each area it was necessary to differentiate levels of performance relative to the school situation. This meant identifying appropriate objectives for each of the four test levels covering grades K through 12 as well as controlling for vocabulary and readability at each level. This was done for both the English and Spanish versions. Within each level the test had to provide for a range of difficulty so that different levels of proficiency could be identified for placement and to permit measurement of gain.

One other major consideration was the average difficulty of the test. Because it is designed to measure the level of English proficiency of students whose English skills are limited, the test had to have sufficient "floor" to yield reliable measures of performance by these students. This meant a test of average difficulty for the limited-English-proficient or "target" students but very easy test for the English-proficient group. The criterion for program eligibility was established by the Aspira Consent Decree to be the 20th percentile for total test raw score on the English proficient norms. It was desirable to maximize the discrimination of the test for the target group at that point, that is, the test should be of 50% average difficulty for the target group at the point represented by the 20th percentile on the English norms. Because it was not possible to identify that raw score point in advance, the test was designed to have an average p-value of about 80 for English-proficient students, and a p-value of 50 for the target group. Because the Spanish-proficient students for whom the test was designed may have had their Spanish contaminated by residence in an English-speaking environment, it was assumed that they would be less proficient in Spanish than would be the English-proficient in English and, therefore, the Spanish version was designed to have a somewhat lower average p-value of about 70 for the Spanish-proficient group.

As in building any test there are constraints. Two that had to be addressed were testing time and response format. For practical reasons the test could not be too time-consuming and also had to be in a group-administered, multiple-choice format to facilitate administration and scoring. The single exception was the measurement of speaking which required from a student a free response to pictorial and/or oral stimuli. Also, it was recognized that "writing" should be measured by means of a writing sample. Because of the very practical problems involved in scoring, however, it was necessary to measure "writing" by means of multiple choice items. As a result the writing subtest basically measures selected elements of language usage that are necessary to good writing. From level to level there is a progression from an emphasis on parts of speech to an emphasis on sentence construction.

The measurement of listening posed the greatest problems. It is desirable to measure the aural comprehension of continuous discourse, not just isolated words, while at the same time minimizing the effect of memory. At the various test levels listening was measured in different ways. The usual

picture vocabulary yielded to the measurement of concepts such as "all, some, both, same, different, etc." presented in the context of a complete sentence and related to pictorial content. Another item type consisted of dictated questions and possible answers involving progressively more complex construction from level-to-level. In order to minimize the effect of memory, the possible answers, but not the questions, were printed in the test booklet. This, of course, introduced the possibility of contamination resulting from the effect of being able to read the options. A somewhat innovative item type consisted of a "listening-cloze" procedure. These items measured continuous discourse in that each consisted of a dictated set of 5 or 6 sentences on a single topic. Each sentence in the set had the last word deleted in order that sentence flow not be disrupted. Again the possible responses were dictated as well as printed in the booklet which represented a trade off: minimize the effect of memory but permit possible contamination from reading. It should be noted that studies are planned to investigate this.

In the measurement of reading a modified cloze approach was adopted in which every Nth word was deleted; the only words deleted were norms, verbs, adjectives and adverbs. A cloze procedure was used because it provides a better means of developing passages of comparable difficulty in two languages since difficulty depends primarily upon the passage itself rather than on a combination of passage difficulty and separate skills item difficulties. Also, with the same amount of reading content, cloze permits a greater number of items which is condusive to higher reliability.

Items were prepared to measure identified objectives with particular attention paid to curriculum relevance, vocabulary level, readability, difficulty of language construction, and bias. Acceptable items were then assigned to pilot test forms. In each language the test consisted of four levels with four subtests each and with three parallel pilot forms at each level. Three student samples were selected. The English version was administered to a sample of English-proficient students and to a sample of limited-English-proficient students, the "target" group. The Spanish version was administered to a sample of Spanish-proficient students. Each sample consisted of 200-400 students per form, per grade. The English-proficient sample were from schools having a large proportion of "monolingual" English-speaking students. The schools were selected so that for any pilot form the mean reading score for the combined selected schools was either plus or minus one month from the New York City mean reading score on a nationally normed test. (The New York City mean reading score on the nationally-normed test was very close to that of the nation as a whole.) The Spanish-proficient sample was derived from schools randomly selected from among those with a large number of students enrolled in Spanish bilingual programs. The identified schools then selected the students to be tested on a random basis. The limited-English-pcoficient or "target" sample who took the English version consisted of all those students who took the Spanish version plus students of other language backgrounds (Chinese, Korean, Greek, etc.) who were in bilingual or ESL programs. About 75% of the target sample were Hispanic which is about their proportion in the population of limited-English-proficient students in the New York City schools. The pilot tests were administered in Fall 1981.

The results of the pilot testing were analyzed using classical as well as Rasch latent-trait procedures. The use of latent-trait procedures was considered particularly appropriate because of the hypothesis of one underlying language proficiency trait. However, data for each subtest and total test were analyzed separately. The pilot results for each language version provided data for item selection for two final parallel forms as well as for the equating of those forms. Pilot data also provided the basis for the development of English-proficient norms. Spanish and limited-English-proficient norms would be derived from administration of the final test, but it would be extremely difficult to have another administration to an English-proficient sample because the time required for its administration would result in an unjustifiable loss of instructional time for students who would never benefit directly from the test.

Item selection for the two parallel forms in each language version was based upon considerations of content and psychometric properties as well as practical considerations of test length. The objectives selected for measurement were the same for each of the two final parallel test forms at each level in each language. The rare exceptions occurred when a particular objective was totally inappropriate for one of the languages. Selecting items based upon their psychometric properties made it possible to construct test levels with an appropriate range of difficulty as well as to construct statistically parallel forms at each level within each language version. Rasch procedures were used in equating forms.

The Spring 1982 final test administration provided the basis for the limited-English-proficient and Spanish-proficient spring norms. Spring norms for the English-proficient group were interpolated. Separate English-proficient and limited-English-proficient norms are required because of the different purposes of the test. Limited-English-proficient students must be compared with English-proficient students in order to know if their performance is such that they can participate effectively in the same instructional situation. Limited-English-proficient norms are also required in order to measure more appropriately gain in performance by these students as they proceed through a program since the limited-English-proficient norm group serves as a "proxy" control group. English-proficient norms would be inappropriate for this purpose. Spanish-proficient norms are required in order to reveal language of dominance for Hispanic students as indicated by their performance relative to the two "proficient" norm groups.

The results of the Spring 1982 and 1983 administration of Forms A and B yielded evidence that LAB-1982 is sensitive to a fall-to-spring and spring-to-spring gain in language proficiency as measured by grade p-values obtained by limited-English-proficient students. Also the results confirmed the fact that the test was of an appropriate level of difficulty for the limited-English-proficient group in that it had sufficient floor to permit reliable measurement of their performance. Grade reliabilities for limited-English-proficient students on the English version for individual subtests ranged from .80 to .95 with 80% above .87, and for total test ranged from .92 to .97. Reliabilities for the Spanish version were comparable. The criterion

cut-off set by the court at the 20th percentile based on English-proficient norms corresponded to percentiles in the 60s based upon limited-English-proficient norms.

Validity is, of course, the most important quality of any test. A test must be valid for a particular purpose and for a specified group. LAB-1982 has a two-fold purpose for a defined group. It must identify students whose English-language proficiency is insufficient for them to participate effectively in an English-speaking instructional environment and it must also be useful for the measurement of gain in English-language proficiency as a result of exposure to an instructional program. These two purposes are closely related since both a placement test and an evaluation test must be sensitive to different levels of English-proficiency and to changes in levels of proficiency. Both must also reflect curriculum and be sensitive to different levels of English proficiency. Curriculum relevance was assured by the curriculum objective to test objective match. Sensitivity to curriculum and to different levels of English proficiency was evidenced by changes in p-value from fall-to-spring and spring-to-spring test administrations. Also tests that serve both placement and evaluation functions for the limited-English-proficient group should be particularly sensitive at and below the criterion cut-off. The results support that this was accomplished. It is a very easy test for the English-proficient but is of average difficulty for the target group. Also it has sufficient "floor" for the target group which makes it suitable for measuring their performance.

The validity of LAB-1982 for placement is based on how appropriately the criterion cut-off score is set. This was mandated by the court as the 20th percentile for total test raw score based on English-proficient norms. The criterion cut-off also effects evaluation because it determines which students enter a program being evaluated. The criterion set by the court appears to work reasonably well but empirical justification is necessary. Such a study is in process and some preliminary work has been done.

Homogeniety of test content is also basic to validity for placement and evaluation purposes. For placement it permits a comparison of the skills of the English-proficient with those of the limited-English-proficient students because it assures that the test focuses on the underlying English proficiency variable. For LAB-1982 this was supported by data from the pilot testing. It was hypothesized that if the test is homogeneous and does, in fact, measure one underlying variable, then it should perform in the same way across the entire language proficiency continuum. That is, item difficulties for the same items should rank order in the same way for English-proficient students as for limited-English proficient students, although there may be considerable differences in levels of difficulty. Rank ordering of item difficulties was found to be extremely similar for these two groups who do have very different levels of English-proficiency skills. Rank-order correlations by subtest within level for Form A ranged from .70 to .93 with 80% above .78 and for Form B ranged from .68 to .96 with 80% also above .78. It should be noted that the lowest rank-order correlations were for the speaking test. The high $KR_{20}$ reliability data by grade are further evidence of LAB's homogeneity.

It is hoped that LAB-1982 subtest scores will be useful for, at least, limited diagnostic purposes. However, until research data are generated including intercorrelations among the subtests, no such claims can be made. Also, data of this type are required to validate "The Partial Divisibility Hypothesis." In the case of Hispanic students, however, it may be productive to compare performance on the English version with that on the Spanish version. There are very different instructional implications for the limited-English- proficient student who reads poorly in both Spanish and English as compared with one who is a highly proficient reader in Spanish but is reading only at a minimum level in English.

To summarize, developing a test to measure the language proficiency of individuals in a language other than their native language introduces unique measurement issues. These are engendered by the two-fold purpose of such a test: program placement and program evaluation. Two different sets of norms are required: language proficient norms and limited-language proficient norms. The test must be constructed so that difficulty level for both groups is such that appropriate placement and evaluation decisions can be made for the limited-English-proficient students. This, of course, implies that the test must be particularly sensitive at and below the criterion cut-off. Concomitant with this consideration is the necessity for the test to focus on the underlying language proficiency variable so that the test performs in the same way across the entire language proficiency continuum. Research conducted in conjunction with the Language Assessment Battery - 1982 supports the conclusion that these issues were sucessfully addressed.

Much research has been conducted in connection with the Language Assessment Battery but further research is in progress.

IE:rt-0954D/35

## REFERENCES

Carroll, John B., 1961, "Fundamental considerations in testing for English proficiency of foreign students". Testing the English Proficiency of foreign students. Washington, D. C.: Center for Applied Linguistics, 31-40.

Language Assessment Battery - 1976. New York: New York City Board of Education, 1976

Oller, John W. 55. 1979. Language tests at school. Longon, England: Longman Group Limited.

Oller, John W., Jr. and Kyle Perkins (ed.). 1980. Research in language testing. Rowley, Mass.: Newbury House.

MMA:rt:
0954D/35-42